In conclusion, therefore, the reference bias shown in this study seems to be real. Such a finding has important implications, since there is no reason to believe that rheumatologists are more biased than others in selecting references. A reader tracing the literature on any new drug using the reference lists given in the articles might risk obtaining a biased sample. Reference bias has another serious implication: it may render the conclusion of the individual article less reliable. Is this also true for review articles, and for other disciplines in medicine?

The study was supported by a grant from the Danish Medical Research Council. I thank the University Library II, Copenhagen, the medical companies, and Alice Nørhede, librarian at Herlev Hospital, for help in data collection; Dr John Anderson for linguistic help; and, especially, Dr Thorkild I A Sørensen, liver unit, Hvidovre Hospital, for his valuable suggestions and comments on the manuscript.

References

1 Poynard T, Conn HO. The retrieval of randomized clinical trials in liver disease from the medical literature: a comparison of MEDLARS and manual methods. *Controlled Clin Trials* 1985;6:271-9.
2 Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Perusing the literature: comparison of MEDLINE searching with a perinatal trials database. *Controlled Clin Trials* 1985;6:306-17.
3 Sackett DL. Bias in analytic research. *J Chronic Dis* 1979;32:51-63.
4 National Library of Medicine. *Medical subject headings: annotated alphabetic list*. Bethesda, Maryland: NLM, 1985.
5 Institute for Scientific Information. *Science citation index. Journal citation reports*. Philadelphia: ISI, 1986.
6 Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chronic Dis* 1967;20:637-48.

Towards a reduction in publication bias

ROBERT G NEWCOMBE

Abstract

Current practice results in the publication of many research studies in medical and related disciplines which may be criticised on the grounds of inadequate sample size and statistical power. Small studies continue to be carried out with little more than a blind hope of showing the desired effect. Nevertheless, papers based on such work are submitted for publication, especially if the results turn out to be statistically significant. There is confusion about what makes a result suitable for publication. Often there is a preference for statistically significant results at the peer review stage. Consequently published reports of small studies tend to contain too many false positive results and to exaggerate the true effects.

The use of a criterion of a posteriori power does not eliminate the bias; a priori power is the criterion of choice. This could be implemented by peer review of study protocols at the planning stage by funding bodies and journals.

Introduction

Profound biological and behavioural differences between human beings mean that statistical methods have to be used in presenting medical research findings in an unbiased way. Hence statisticians have devised methods of estimation and significance testing, which are now widely used. Nevertheless, though the mathematical aspects of these methods are acceptable, what is done with the results commonly leads to serious selection bias. An article that reports a statistically significant difference between two treatments is more likely to be published than one which does not. Many research studies have inadequate numbers of subjects, and significance can be attained only if chance conveniently exaggerates the difference.

So long as statistical significance is used as a major criterion of acceptability for publication the published results of medical research will contain a high proportion of false positive results. Thus quantitative estimates of treatment effects taken from published work cannot be regarded as free from bias. There are established methods to calculate the power of a study, which is the probability of detecting a specified, important difference using a test with a set significance level. The interpretation of statistical power is satisfactory only when it is calculated with values specified at the design stage of the study. The proper method to assess the adequacy of the sample size is by peer review of values specified in the protocol. If this is done the significance level eventually attained is no longer relevant to selection for publication.

Importance of sample size

Manuscripts submitted to medical journals often contain serious statistical faults.[1] Various steps have been taken to remedy this,

Department of Medical Computing and Statistics, University of Wales College of Medicine, Cardiff CF4 4XN
ROBERT G NEWCOMBE, MA, PHD, lecturer in medical statistics

notably the checklists used by the *BMJ*,[2] and there is now also an increased awareness of the need for therapeutic efficacy to be evaluated with randomised controlled trials. Nevertheless, power calculations are still rarely used.[3]

Conventional significance testing (table I) leads to great emphasis on the type I error rate α, but the type II error rate β and its complement, the power $1-\beta$, though very important, are neglected.[3] In particular, in a clinical trial the number of subjects required depends on the α and β levels chosen, the treatment difference of interest, and the degree to which the treatment effect varies between subjects. The choice of the first three of these is somewhat arbitrary, and the fourth may be difficult to estimate. Nevertheless, the study is likely to be valid only if values are chosen for these parameters and the resulting sample size requirement determined, whether by the use of formulas,[4] diagrams,[5] or tables.[6]

TABLE I—*The significance testing paradigm. α and β denote the frequencies with which type I errors and type II errors are made. $1-\beta$ is known as the power of the test*

	Decision from significance test	
True state of nature	Accept H_0	Reject H_0
H_0 valid	True negative $1-\alpha$	False positive α
H_0 not valid	False negative β	True positive $1-\beta$

The most obvious consequence of an inadequate sample size is that investigators may well not show a clinically important effect. Such a false negative result, if propagated by publication, is apt to be widely misinterpreted as a demonstration that there is no difference between the treatments. This has provoked two responses among those who decide what is to be published. Firstly, statisticians advocate a shift of emphasis away from significance testing and towards estimation and confidence intervals.[7] A wide confidence interval is understood as implying that large, potentially important differences cannot be ruled out. The confidence interval approach may also help in a wider context—for instance, in showing that the results of two apparently disparate studies are not incompatible, the truth perhaps being somewhere between their two estimates.

The second response is to exclude small studies, with high β, from publication. There are three approaches in which this may be done. Firstly, attainment of a desired level of significance may be used as a criterion. This seems plausible because, for a fixed α, both the attained significance level p and the type II error rate β reflect the sample size. Nevertheless, p (unlike α) depends on sampling variation and the use of this criterion leads to publication bias. Secondly, assessment based on statistical power calculated from the data gives the appearance of greater soundness; it does not fall into the obvious trap of the first approach and is based on data rather than on the uncertainty of a prior targeted difference. In reality, however, the requirement of significance using an α level of 0·05 and an a posteriori β of 0·2 amounts to nothing more than statistical significance at a more stringent level of $\alpha=0{\cdot}005$ and thus also does not avoid publication bias. This kind of β value is akin to p, not to α, and includes sampling variation. The third approach is to impose the requirement of an adequately low β value assessed a priori; this does not lead to bias since the β value is not subject to sampling variation.

Thus results based on studies which had a poor prospect of yielding useful information may justifiably be rejected, but only if the criterion is based on power assessed a priori.

Nature and consequences of publication bias

Publication bias may be defined simply: significant results are preferred for publication. Attention was drawn to it as early as 1963[8] and it has been "rediscovered" several times since. Suppose the α rate chosen is 0·05. Then, just 5% of studies in which H_0 is valid will yield a test statistic significant at the 5% level. If attention is limited to studies that attain publication, however, the proportion of such false positive results is higher. The significance testing paradigm does not permit us to say what proportion of statistically significant results are false positives, but the effect of publication bias is to make this proportion disquietingly larger than it would otherwise be.

Correspondingly, studies selected for publication tend to contain exaggerated estimates of the main effects, and trials with truly modest treatment effects will achieve statistical significance only if random variation conveniently exaggerates these effects.[9] Conversely, variation is underestimated. These biases operate more strongly the more inadequate the sample size. A study with low power, where the true treatment effect is zero or small, must grossly exaggerate it (by chance) to show significance and attain a prospect of publication. False positives and exaggerated estimates may well dominate much of medical publication. This phenomenon is likely to contribute to the disparity commonly found in the results of different studies, which leads to controversy instead of well established, consistent findings. The desire to minimise the impact of false positive assertions may result in a preference for publishing findings which refute a previous claim, rather than confirmatory results—a further source of bias.

Such selection bias may equally be introduced by the editorial team (editorial selection bias) or by the researcher or supervisor or head of department (submission selection bias). At each stage a significant result may be construed as particularly encouraging and failure to attain significance as correspondingly discouraging. This operates in addition to any biases introduced because of prejudice.[10]

Publication bias continues to arise only because two conditions hold: the criteria for selecting studies for publication are inadequate, and many studies performed and submitted for publication have been done on small numbers of subjects. Significance testing, the time honoured framework for inductive inference, is evidently deficient as a selection criterion. Nevertheless, the confidence interval approach incurs the same danger of publication bias: studies in which the confidence interval for the size of the effect excludes zero are likely to be preferred for publication—a condition that is equivalent to statistical significance. It has been asserted that overconcentration on simplistic significance testing is responsible for most of the ill based criticisms of small trials.[11] The more careful approach using confidence intervals overcomes many of the difficulties. But so long as confusion remains as to what constitutes a result warranting publication a bias will ensue from submission and editorial selection processes.

The other prerequisite for publication bias is the widespread use of inadequate sample sizes. The other consequence of this is that a doctor seeking information to guide a clinical decision is confronted with a bewildering variety of conflicting claims. To remedy this dilemma "meta-analyses" or "overviews" have been constructed, which fit together results of several studies and seek to make the best use of data from studies which would otherwise yield little information. Nevertheless, published studies are still a biased sample of all the relevant work that has been done. The only prospect of eliminating this bias is to contact all investigators who may have done relevant work and ask for their unpublished data. Iain Chalmers and Thomas Chalmers are pursuing this goal in connection with the Oxford Database of Perinatal Trials, and their work should provide some evidence on the quantity of "negative" studies that either never get written up or never get published.

The high prevalence of small studies stems from the way that research is organised. Much material submitted for publication has come from studies that are regarded as the work of an individual researcher, performed within severe constraints of time and resources; often there is little more than a blind hope that the desired effect will be shown. Research output remains a major criterion for assessing candidates for promotion and so on, even though it is widely recognised to be deficient. When research output is equated with publication, however, the consequences for the standards of published work are grave. The constraints an individual investigator faces often preclude obtaining results of

external validity, but publication in a highly regarded, widely circulated journal implies such validity, however mistaken this is given the background of inadequate statistical power.

Thus the researcher faces a dilemma: on the one hand, most studies he can perform will need the collaboration of others to attain adequate statistical power; on the other hand, any collaborative study (even if it is feasible) will deprive him of personal kudos. Only those who are remote from the researcher's dilemma—journal editors and referees, funding bodies, and (to a lesser degree) ethical committees—can uphold the highest scientific standards with no conflict of loyalties. These agents are not obliged to accept the status quo and can refuse to support or publish inadequate research. I regard it as their prerogative, if not obligation, to do so.

A radical proposal

Selection of work for funding or publication, then, should primarily be based on reasonableness a priori: Has the design adopted (explicitly or implicitly) a good prospect of yielding useful information? "Design" here includes the study idea, scientific basis, clinical relevance, originality, and so on, as well as the study's structure and the number of subjects. If all this is satisfied then the paper should be published irrespective of whether statistical significance or the targeted size of difference was attained. The difference actually observed is irrelevant to the decision (see Mahoney,[10] p 163). The assessment of scientific validity would therefore be the same, whether carried out before the study or after it. The only additional requirement a posteriori is adequate adherence to the protocol—in particular, attainment of the planned sample size.

The consequences of this shift in emphasis to a priori criteria are most important in the case of studies of inadequate power. Table II contrasts what would happen to the results of these studies under the proposed rule with what is likely to happen at present. The publication of "positive" findings would be inhibited. The advantage would be the exclusion of false positives from inadequate studies, with their grossly exaggerated estimates of differences. Against this must be weighed the cost of failing to publish true positives—which would occur quite often ($1-\beta=0{\cdot}5$), but which are based on inadequate evidence and also overestimate the difference.

Application of this principle to studies with adequate power would lead to more widespread publication of negative results (table III). True negative results would be salvaged from studies of acceptable power—though these might currently be accepted anyway, especially if supplemented with confidence intervals. This

TABLE II—*Consequences of a shift to assessment by a priori power. The case of a study with inadequate power: $1-\beta=0{\cdot}5$, $\alpha=0{\cdot}05$*

	Decision from significance test	
	Accept H_0	Reject H_0
Whether published under:		
Present practice	?No	Yes
Proposed policy	No	No

would be at the cost of publishing studies with false negative results, though these would not be too frequent ($\beta=0{\cdot}1$).

Both journal editors and funding bodies can and should require specification of statistical power. They should require that a protocol or a write up should describe clearly the details of the design of the study—in particular, the following:

(*a*) the structure;

(*b*) the choice of the most appropriate criterion variable on which to base the power calculation and the most appropriate groups to be compared;

(*c*) the size of the effect to be reliably detected and (except in the case of a binary variable) how much this effect varies between subjects;

(*d*) the sample size (specifying accrual rate and period) aimed at, with specific allowance for expected dropouts;

(*e*) consequent statistical power and the method by which it was derived.

These parameters should be identical in the protocol and in the eventual study report. The same criterion should be used to assess validity at both stages—in particular, the write up should be assessed on the basis of the values laid down before any data were collected. The only additional requirements at the publication stage would be the completion of the study as laid down in the protocol, with full information on as many subjects as were contracted for; variability in response between subjects not grossly in excess of that planned for; and the usual standards of adequate analysis, inference, and discussion.

TABLE III—*Consequences of a shift to assessment by a priori power. The case of a study with adequate power: $1-\beta=0{\cdot}9$, $\alpha=0{\cdot}05$*

	Decision from significance test	
	Accept H_0	Reject H_0
Whether published under:		
Present practice	?No	Yes
Proposed policy	Yes	Yes

This approach entails assessment of the parameters assumed on an a priori basis; they are to be judged in the light of knowledge current at the time the study was designed. Other results coming to light during the study should not be allowed to affect the judgment of validity (though occasionally a major advance occurring during this period may render the results no longer relevant).

Journal editors as well as grant awarding bodies could implement this proposal most effectively by requiring submission of protocols for peer review at the planning stage. In either case an independent review body could be used. Specialists in the subject could assess the reasonableness of the values supplied for the parameters on which the power calculation is based (particularly the smallest clinically important difference), and the verification of the power calculation would not be a formidable task for a statistician or other assessor familiar with this. These assessments, once performed for the protocol, would not need to be repeated for the write up. Consequently, having accepted a protocol as adequate and relevant, a journal could offer eventual publication, conditional only on completion of the study in adequate conformity to the protocol together with the usual requirements of adequate analysis, inference, and discussion. It would become normal practice to accept an article only if this had been done.

The work of Mahoney suggests that reviewers may find it difficult to comment on incomplete manuscripts.[10] Nevertheless, Mahoney's study is not an ideal model for the process I advocate, for two reasons. Firstly, his reason for the incompleteness of the manuscript was inadequate. It would be understood, however, that the material to be evaluated was only a protocol, even though it would be virtually unaltered in the eventual article—and this would become an accepted element of peer review (as it is, to a limited extent, with funding bodies). Secondly, Mahoney studied psychologists known to have entrenched, diametrically opposite beliefs, to a degree (I hope) not encountered often among doctors; knowing that results would shortly be disclosed, they would be reluctant to commit themselves unequivocally to a favourable stance, lest the results turned out to contradict their chosen position. At the stage of review of a protocol this possibility is more remote.

To put these recommendations into practice would be more feasible for formal, well structured study designs, such as the clinical trial, than for less formal explanatory work—for which the rationale of significance testing is more contentious. Like other alterations in editorial policy, this would best be introduced as a decisive change, as from a given date, with advance indication given, as a piecemeal approach to change is unlikely to work.[12] I

hope that enlightened editors will take up the challenge; the lead must come from an established, prestigious journal that can afford to be choosy.

Conclusion

Publication bias is endemic and will remain so as long as the sample sizes commonly used in research are too small and the methods used to assess adequacy of sample size are deficient. Assessment by a priori criteria—in particular, systematic peer review at the planning stage—would result in a much tighter measure of control over the quality of published work, with the prospect of improvement in study design in general and statistical power in particular.

I thank several colleagues, especially Dr Edward C Coles and the *BMJ* editorial team and the referee, for constructive comments.

References

1 Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA* 1966;195:1123-8.
2 Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. *Br Med J* 1986;292:810-2.
3 Freiman JA, Chalmers TC, Smith H, *et al*. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978;299:690-4.
4 Armitage P. *Statistical methods in medical research*. Oxford: Blackwell, 1971:186.
5 Altman DG. Statistics and ethics in medical research: III. How large a sample? *Br Med J* 1980;281:1336-8.
6 Cochran WG, Cox GM. *Experimental designs*. 2nd ed. New York: Wiley, 1957:24-5.
7 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;292:746-50.
8 Melton AW. Editorial. *J Exp Psychology* 1962;64:553-7.
9 Pocock SJ. Current issues in the design and interpretation of clinical trials. *Br Med J* 1985;290:39-42.
10 Mahoney MJ. Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research* 1977;1:161-75.
11 Powell-Tuck J, MacRae KD, Healy MJR, Leonard-Jones JE, Parkins RA. A defence of the small clinical trial: evaluation of three gastroenterological studies. *Br Med J* 1986;292:599-602.
12 Lock S. *A difficult balance—editorial peer review in medicine*. London: Nuffield Provincial Hospitals Trust, 1985.

(Accepted 11 June 1987)

Medicine and the Media

AT THE ANNUAL scientific meeting of the British Paediatric Association last year the prize for the best paper presented by a young paediatrician went to a member of a research group from Oxford. Papers offered for the annual meeting are examined by the association's academic board not only for their scientific worth but also for adherence to ethical standards. This paper, later published in the *Lancet*,[1] has now been condemned by certain sections of the press and by a group of members of parliament. What was the work so condemned?

Preterm infants of low birth weight live at considerable risk, particularly of cardiorespiratory failure, and the risk is increased if they have to undergo an operation. Clinical experience suggested that deep anaesthesia and narcotic analgesics would increase the risk. That and the belief that such infants have a poor perception of pain because of lack of myelinisation in the central nervous system led to the conventional practice of anaesthesia with nitrous oxide and muscle relaxants combined with artificial ventilation. In a study of 40 published reports the Oxford team found that three quarters of newborn babies undergoing surgical ligation of patent ductus arteriosus had received muscle relaxants alone or with nitrous oxide.

In the preterm infant with a poor or absent ability to cry it is difficult to tell clinically whether pain and stress are being experienced, but newer biochemical methods that detect hormones and intermediary metabolites associated with stress now make the assessment of stress more possible and prompted a re-examination of the problem by the Oxford team. The team wanted to find out whether adding a little narcotic analgesic to the accepted anaesthetic regimen might prove beneficial rather than harmful. Using these metabolic methods, they therefore compared the response to surgical ligation of patent ductus arteriosus carried out under the conventional regimen with and without the narcotic analgesic fentanyl. The possibility that fentanyl might adversely affect respiration and circulation postoperatively was also studied.

A randomised trial was designed with help from the National Perinatal Epidemiology Unit in Oxford to ensure that the results were statistically valid and that a meaningful result would be recognised as soon as possible. After only eight babies in each group had been operated on the results showed that the new regimen was significantly superior to the old not only in reducing the stress response estimated biochemically but also in improving the postoperative state. Thus for the first time good scientific evidence was produced of the need to provide deeper anaesthesia during operations on these tiny infants.

This research was commended by the distinguished American paediatrician Dr William Silverman, author of the widely acclaimed book *Human Experimentation: A Guided Step Into the Unknown*.[2] He wrote that the Oxford workers "deserve a loud vote of thanks for the ethically sound effort to subject to a rigorous test opinion based on long standing practice. And their call for further study should not fall on deaf ears. It is indeed urgent to determine the pathophysiological consequences of unrelieved pain and suffering inflicted during everyday care of newborn babies."

Members of the British Paediatric Association were thus amazed and the doctors who had done the work bewildered and distressed when after a distorted report in the *Daily Mail* entitled, "Pain-killer shock in babies' operations" (8 July) this work became the subject of a condemnatory "press release: for immediate publication" issued by some members of parliament forming the All Party Parliamentary Pro-Life Group. The *Lancet* article appeared in January, the story in the *Daily Mail* in July, and the press release from the members of parliament in August. The press release was entitled "Inhumane baby operations slammed" and the first paragraph stated:

"Fourteen members of parliament have demanded an inquiry into trials in which sixteen premature babies were given open heart surgery, eight of them without the use of pain killers to test whether or not the babies could experience pain."

The press release then said that the General Medical Council was being asked to investigate these trials with a view to bringing those responsible before its disciplinary committee. It continued:

"In a statement Sir Bernard Braine said:

'The trials seemed to us to be even more barbarous when one considers that the babies being tested for pain were given curare, a paralysing drug, so that they would have been unable to kick or struggle even if they were in agony, the obvious intention being to keep them immobile at all costs throughout the operation. Apart from this they were given only nitrous oxide (laughing gas).'"

Implying misleadingly that wisdom acquired from the research existed before it was carried out the statement went on:

"Not surprisingly post-operatively they fared far worse than the eight babies who were given pain killers. Two of the disadvantaged babies suffered from hypotension, two showed poor peripheral circulation—both of which can be indications of shock which most